BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 8 1998
Pages 751–752

PPCMatrix: a PowerPC dotmatrix program to compare large genomic sequences against protein sequences

Thomas R. Bürglin

Department of Cell Biology, Biozentrum, University of Basel, Klingelbergstrasse 70,CH-4056 Basel, Switzerland

Received on May 26, 1998; revised on June 24, 1998; accepted on June 25, 1998

Abstract

Summary: *An interactive dotmatrix program for the MacOS was designed that allows comparison of DNA to protein sequences using nested 3-frame translations.*

Availability: *Shareware, available at http://copan.bioz.unibas.ch/software/*

Contact: *burglin@ubaclu.unibas.ch*

Implementation and discussion

An effective technique to compare sequences interactively is the dotmatrix procedure (Staden, 1982). It is not only useful for searching small regions of similarity, for example, regulatory regions of genes, but also for comparing sequences over larger distances. Recently, much genomic sequence material has become available, for example, about 79% of the *C. elegans* genome is now sequenced (Waterston *et al*., 1997). The open reading frames predicted by computer methods, such as Genefinder within ACeDB (Durbin and Thierry Mieg, 1991–), can be wrong or erroneous, and manual inspection is necessary to match regions of similarity to open reading frames and to join the correct exons. Sonnhammer and Durbin (1995) developed a dotmatrix program for UNIX machines that can display plots of translations of all reading frames against a protein sequence in a single plot to aid in such analysis. In an independent approach, a program was designed that can handle small as well as large DNA and protein sequences and the DNA sequences can be compared to protein sequences after translation in three frames. The results are displayed in a color dotmatrix, where color indicates the frame. The software was written in Procedural C for the MacOS and compiled for PowerPC as well as 68k processors.

The basic algorithm is based on a previous version of the program written for the ATARI ST and PC compatible computers using the GEM windowing system (Bürglin and Blumberg, 1987, unpublished). New scores are added to the scan window and old ones subtracted, rather than integrate each time over the whole scan window. However, rather than assigning a memory variable for each scan window of the X-axis (as in Karreman, 1992), only one variable is used to store the value integrated over the window length; the calculation of this variable progresses in a diagonal fashion across the plot, rather than horizontal, as in Karreman (1992). Thus no memory is necessary for storing all the X-axis scores, allowing unlimited sequence lengths. The drawback is that within the innermost program loop for each matrix point the coordinates for both axis need to be incremented. To reduce the calculation time as much as possible, the boundary conditions are calculated for each diagonal. Furthermore, the actual plotting routine is coded several times in slightly modified form to accommodate options.

The major feature of this program is that genomic sequences can be translated in the 3 reading frames and then be compared to a protein sequence. The following principle is applied: the DNA is translated in all three reading frames and stored in a nested fashion: position 1: residue 1 of frame 1; position 2: residue 1 of frame 2; position 3: residue 1 of frame 3; position 4: residue 2 of frame 1, etc. This artificial sequence is now compared to a protein sequence using the dotmatrix algorithm. This produces three different dotplot layers on top of each other. To distinguish the layers of the three frames, each frame is color-coded with a different color.

The matrix is not stored in memory, thus only space for sequences and plot windows (if G-World screen buffers are active) needs to be allocated. Thus, even if only 3 Mb of RAM is allocated to the program, two 1 Million residue long sequences could be compared. This feature is useful, if large genomes are to be loaded, but only selected sub-areas need to be analyzed. The additional memory requirement for the G-World screen buffers depends on color depth screen size and number of windows.

Sequences can be manipulated in many ways, i.e. Translation and Reverse Complement. Sequences can be edited or newly entered in a separate window (with speaking capabilities). Currently recognized sequences types are plain text (including Staden contig marks), GCG and PIR; the sequence length is only limited by the amount of memory allocated to the application. The Dotmatrix window allows

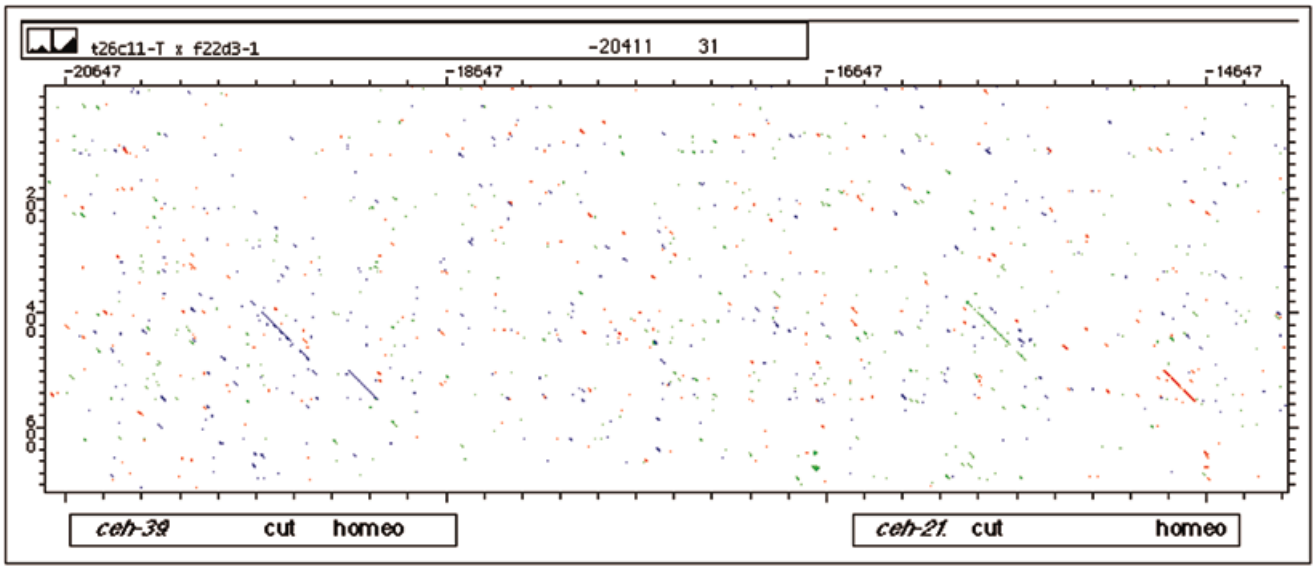

Fig. 1. Screen shot of the dotmatrix window of a 3-frame nested translated *C.elegans* cosmid sequence T26C11 (Genbank U41017) compared to the ONECUT class protein F22D3.1 (on cosmid F22D3, Genbank U28993). The different colors show the matches in different reading frames. Horizontal offsets are due to introns. The cut domain and the homeodomain of the ONECUT genes *ceh-39* and *ceh-21* are indicated.

scrolling and zooming. Interesting regions can be selected to display an alignment.

Several dialogues allow customization of various options, some of which can be stored in a Preferences file. Several matrices for DNA and protein are available. The matrices, as well as the translation tables are stored in two text files and can be modified for special applications.

The plotting times compare favorably with the commonly used GCG program set Compare and Dotplot (Devereux *et al*., 1984). On a DEC 7000-620 AXP under VMS V6.2 using GCG version 9.0, the calculation time of Compare (10 000 by 10 000 residues, window length 21, stringency 14) was 25 s and Dotplot took another 20 s to generate the plot on the screen of a Macintosh using Telnet V2.6 over an Ethernet connection. The PPCMatrix program using the same sequence and stringency settings took 6 s on a PowerMacintosh 8600/250 (16 million points/s). PPCMatrix is also 2.5 times faster than the less flexible dotplot of the commercial software GeneJockey II.

The 3-frame nested translation dotmatrix is especially useful in those cases, when open reading frames in the genomic sequence are interrupted by introns or frameshifts (sequencing errors) and coding regions are found in different frames. The new feature of the 3-frame nested translation was employed to examine the genomic regions of CUT superclass homeobox genes in *C. elegans* (Figure 1). These genes distinguish themselves from the *Drosophila cut* gene, because they have only one cut domain (Lannoy *et al*., 1998). The analysis confirmed that no additional cut domains are present in these novel cut class genes.

Acknowledgments

T.R.B. is supported by START fellowship NF. 3130–038786.93 from the Swiss National Science Foundtion.

References

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res*., **12**, 387–395.

Durbin,R. and Thierry Mieg,J. (1991–) A C. elegans database. Code and data available from anonymous FTP servers lirmm.lirmm.fr, ftp.sanger.ac.uk, and ncbi.nlm.nih.gov.

Karreman,C. (1992) A dotplot program for the Atari ST, for the analysis of DNA and protein sequences. *Comput. Appl. Biosci*., **8**, 75–77.

Lannoy,V.J., Bürglin,T.R., Rousseau, G.G. and Lemaigre,F.P. (1998) Isoforms of hepatocyte nuclear factor-6 differ in DNA-binding properties, contain a bifunctional homeodomain, and define the new ONECUT class of homeodomain proteins. *J. Biol. Chem*., **273**, 13552–13562.

Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.

Staden,R. (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucl. Acids Res*., **10**, 2951–2961.

Waterston,R.H., Sulston,J.E. and Coulson,A.R. (1997) The genome. In Riddle,D., Blumenthal,T., Meyer,B. and Priess,J. (eds), *C. elegans* II..Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 23–45.